БОЧАРОВ Б.П. ВОЕВОДИНА М.Ю.

СТАТИСТИЧЕСКИЙ АНАЛИЗ ПРОПУСКНОЙ СПОСОБНОСТИ INTERNET

За последние годы автоматизация прочно утвердилась в жизни библиотек. Организация электронных каталогов и картотек, составление бюллетеней и библиографических указателей, формирование учебных материалов и их распространение – вот далеко не полный перечень задач, которые возникли и уже решаются в современной библиотеке. Кроме того, назрела необходимость обмена информацией через INTERNET как с коллегами, так и с читателями.

Многие библиотеки выставляют на INTERNET - сайтах свою продукцию. В связи с этим встает вопрос о её дальнейшем использовании. Для того, чтобы наша работа была востребована, нам следует позаботиться об этом.

Материалы на сайте должны быть не только содержательны, структурированы, красиво и наглядно оформлены. Не менее важно сделать так, чтобы заинтересованный пользователь мог успешно скачать их без особых усилий.

Нередко плохое качество INTERNET да и низкая квалификация пользователей не дает возможности скачать достаточно большой файл целиком. Успех этого действия находится в обратной зависимости от размеров скачиваемого файла: чем больше объем файла, тем труднее его получить. С другой стороны, искусственно уменьшая объем файлов, мы рискуем утратить к ним интерес пользователей из -за их малой информативности.

В этой связи хотелось бы реально оценить оптимальные размеры файла, размещаемого на INTERNET-сайте. Для анализа использован известный читателям журнала архивный файл с демонстрационной версией программы «Картотека книгообеспеченности», размещенный на сайте журнала в 2001 году.

Размер файла составляет 3,5 мегабайт.

В качестве метода исследования выбран статистический анализ данных, которые фиксировались с октября 2001 г. по сентябрь 2006 г. За указанный период зафиксировано 6585 обращений к файлу. Статистика нашего сервера позволяет определять IP-адрес пользователя и объем переданной информации. Можно также определить, использовалась ли одна из специальных программ (их называют программами «докачки»), позволяющая скачивать файл небольшими частями.

Следует отметить, что сам по себе статистический анализ не дает ответа вопрос, какой должен быть оптимальный размер INTERNET - страницы; он позволяет лишь оценить вероятность скачивания файла в зависимости от его размера. На основании этой информации, а также руководствуясь собственным опытом, оптимальный размер файла должен определять человек.

Классификация пользователей

На основании статистической информации можно классифицировать (может быть, приблизительно) пользователей по качеству их подключений к INTERNET. Условно мы разделили всех пользователей на три категории.

I. Пользователи, которые не использовали программы докачки и не смогли скачать файл до конца. Качество INTERNET у таких пользователей считаем плохим.

II. Пользователи, которые смогли скачать тестовый файл с использованием программ докачки. Качество INTERNET в данном случае не имеет значения, так как квалификация пользователя позволяет нормально работать даже с низким качеством связи.

III. Пользователи, которые смогли скачать тестовый файл за один раз и, будем считать, с хорошим качеством INTERNET.

Поквартальные статистические данные по каждой категории пользователей приведены в табл. 1.

Табл. 1

Год / Квартал	Количество пользователей			Соотношение в %		
	I	II	III	I	II	III
2001/4	22	9	23	40.7	16.7	42.6
2002/1	13	2	15	43.3	6.7	50.0
2002/2	27	6	22	49.1	10.9	40.0
2002/3	6	1	14	28.6	4.8	66.7
2002/4	20	5	21	43.5	10.9	45.7
2003/1	19	3	19	46.3	7.3	46.3
2003/2	31	5	22	53.4	8.6	37.9
2003/3	12	1	16	41.4	3.4	55.2
2003/4	20	5	23	41.7	10.4	47.9
2004/1	24	4	20	50.0	8.3	41.7
2004/2	22	5	28	40.0	9.1	50.9
2004/3	10	4	20	29.4	11.8	58.8
2004/4	24	5	23	46.2	9.6	44.2
2005/1	21	4	27	40.4	7.7	51.9
2005/2	23	6	28	40.4	10.5	49.1
2005/3	7	1	16	29.2	4.2	66.7
2005/4	26	5	26	45.6	8.8	45.6
2006/1	24	8	28	40.0	13.3	46.7
2006/2	17	6	29	32.7	11.5	55.8
2006/3	9	7	20	25.0	19.4	55.6

Проанализируем данные из таблицы 1. За весь период наблюдений число пользователей составило:

I - ой категории - 377 (41,4 %);
II - ой категории - 92 (10,1 %);
III - ей категории - 440 (48,5 %).

На рис. 1 показано общее количество пользователей (всех категорий), которые скачивали демонстрационную версию картотеки в каждом квартале.

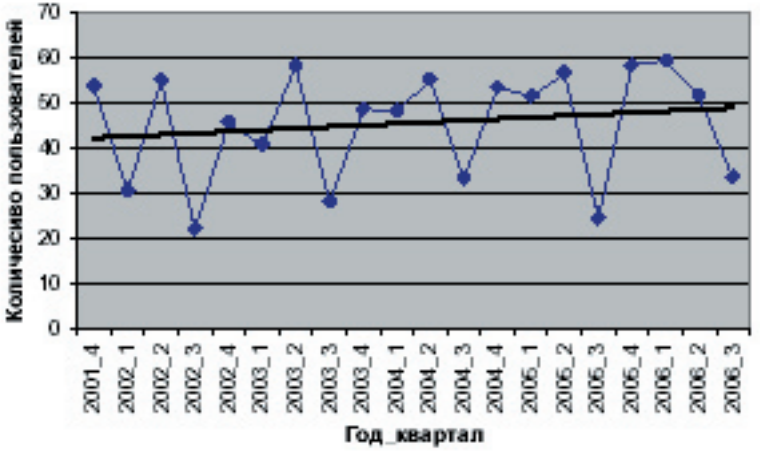


Рис.1.

Прямая линия показывает тренд (тенденцию изменения). Уравнение линии тренда y=0,39x+41,4. Это значит, что общее количество пользователей увеличивается приблизительно на 3,2 % в год (0,8 % в квартал).

Перейдем к рассмотрению процентного соотношения каждой категории пользователей. Начнем с пользователей, у которых плохой INTERNET (см. рис. 2). Уравнение линии тренда

$$y=- 0,54x + 46,0$$

показывает, что число таких пользователей снижается приблизительно на 6,3% в год.

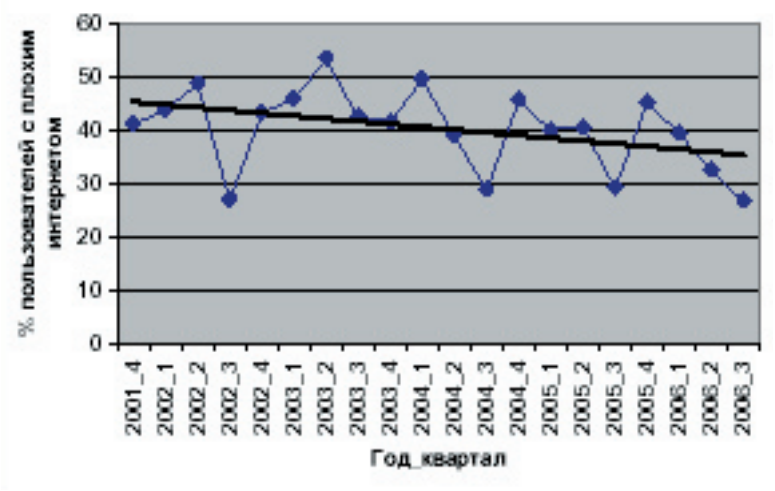


Рис. 2

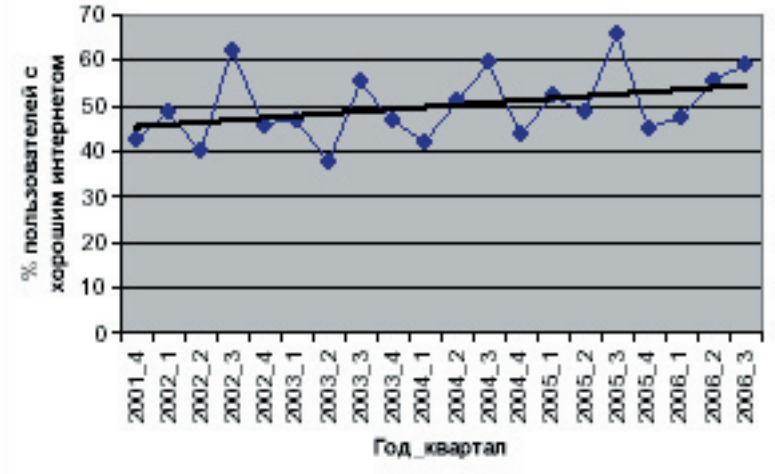


Рис. 3

Процентное соотношение пользователей, имеющих хороший INTERNET в каждом квартале исследуемого периода представлено на рис. 3.

Уравнение линии тренда **y = 0,39x+46,0** показывает, что число таких пользователей увеличивается приблизительно на 2,8% в год.

Рассмотрим теперь пользователей, которые большие файлы скачивают по частям, с помощью специальных программ «докачки» (см. рис. 4)

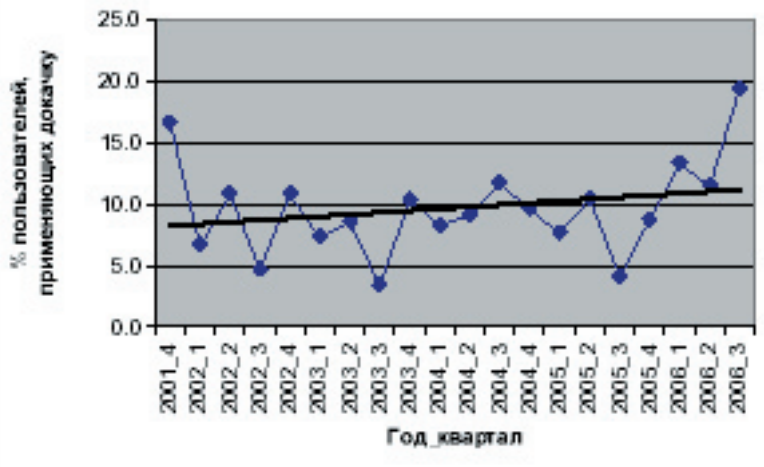


Рис. 4.

Уравнение линии тренда **y = 0,16x+8,1** показывает, что число таких пользователей увеличивается приблизительно на 5,5% в год.

Из всего вышесказанного можно сделать следующие выводы.

1. Наши статистические данные опровергают тезис о стремительном улучшении качества INTERNET. Если существующие тенденции сохранятся, то процент пользователей с плохим интернетом станет равным 0 только в 2023 году. Это дает серьезные основания надеяться на то, что наши исследования будут актуальными еще несколько лет.

2. Статистический анализ в данном случае не позволяет однозначно формально оценить процентные соотношения пользователей различных категорий. Из многих вариантов мы выбрали достаточно консервативный – средние значения за весь период наблюдений. Таким образом, мы считаем, что доли пользователей категорий I, II, III составляют соответственно 0,41, 0,10 и 0,49.

Теперь приступим к следующему этапу анализа.

Анализ «плохого» INTERNET

На данном этапе исследований определим вероятность скачивания тестируемого файла пользователями с плохим качеством INTERNET. Мы выбрали данные для исследования следующим образом. Длина тестирумого файла – 3,5 МБ. При каждом прерывании процесса скачивания файла фиксировалось значение случайной величины (с.в.) **X**, численно равное длине той части файла, которую удалось скачать (в байтах). Объем полученной выборки – 910 элементов.

Интуиция подсказывает, что чем большая часть файла получена пользователем, тем больше вероятность обрыва связи. В технике рассматривается похожая величина, которая называется «наработка на отказ» (например, время работы электрической лампочки). Это – случайная величина, которая распределена по экспоненциальному закону. Проверим интуицию математикой и произведем стандартную процедуру статистической обработки данных.

Для исследования закона распределения с.в. **X** весь интервал, в котором лежат значения с.в., разобъём на 20 равных частей. Границы интервалов разбиения представлены в таблице 2.

Далее определяется количество попаданий с.в. в каждый из интервалов разбиения и соответствующие частоты.

Табл. 2

№	Границы интервалов		Кол-во попаданий в инт.	Частота попаданий в инт.	Плотность распределения (теор.)	Значение критерия C^2
	мин.	макс.				
1	0	183 500	506	0,5560	0,3626	0,1032
2	183 500	367 000	182	0,2000	0,2300	0,0039
3	367 000	550 500	48	0,0527	0,1459	0,0594
4	550 500	734 000	27	0,0297	0,0925	0,0427
5	734 000	917 500	36	0,0396	0,0587	0,0062
6	917 500	1 101 000	13	0,0143	0,0372	0,0141
7	1 101 000	1 284 500	16	0,0176	0,0236	0,0015
8	1 284 500	1 468 000	14	0,0154	0,0150	0,0000
9	1 468 000	1 651 500	6	0,0066	0,0095	0,0009
10	1 651 500	1 835 000	7	0,0077	0,0060	0,0005
11	1 835 000	2 018 500	12	0,0132	0,0038	0,0230
12	2 018 500	2 202 000	8	0,0088	0,0024	0,0167
13	2 202 000	2 385 500	6	0,0066	0,0015	0,0166

14	2 385 500	2 569 000	6	0,0066	0,0010	0,0324
15	2 569 000	2 752 500	4	0,0044	0,0006	0,0231
16	2 752 500	2 936 000	1	0,0011	0,0004	0,0013
17	2 936 000	3 119 500	5	0,0055	0,0002	0,1106
18	3 119 500	3 303 000	6	0,0066	0,0002	0,2625
19	3 303 000	3 486 500	4	0,0044	0,0001	0,1844
20	3 486 500	3 670 000	3	0,0033	0,0001	0,1647
	Сумма		**910**	**1,0000**	**0,9913**	**1,0678**

Вид распределения эмпирических частот позволяет выдвинуть гипотезу об экспоненциальном законе распределении с.в. **X** со средним выборки $\lambda = 403042$.

Для проверки выдвинутой гипотезы табулируем теоретическую плотность распределения в точках, являющихся серединами интервалов разбиения и вычисляем значение критерия χ^2.

Значение критерия $\chi^2 = 1{,}0678$ не входит в критическую область, следовательно, гипотеза об экспоненциальном законе распределения с.в. **X** принимается (естественно, с вероятностью не равной 1; в нашем случае, с вероятностью 0,95).

Рис. 5 дает возможность сравнить визуально эмпирическую частоту и теоретическую плотность распределения ***f(x)*** исследуемой с.в.

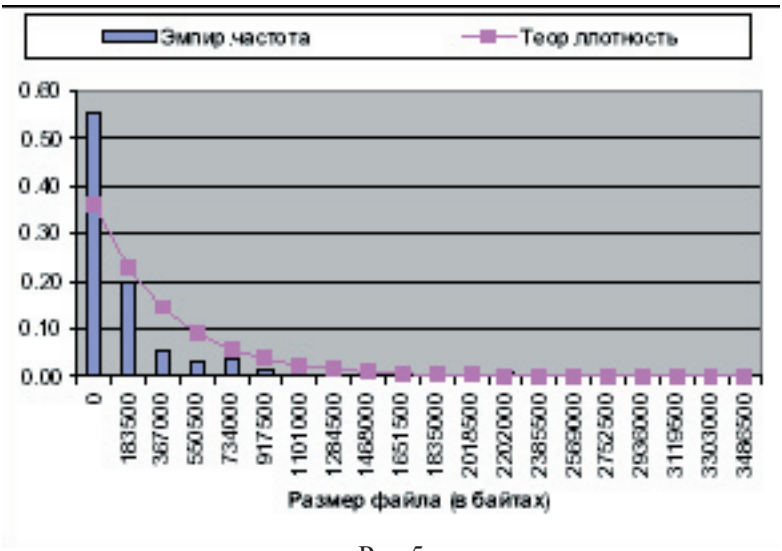


Рис 5.

Построение теоретического закона распределения в виде плотности ***f(x)*** с.в. **X** позволят вычислить вероятность скачивания файла пользователем INTERNET плохого качества в зависимости от его длины. В нашем случае «плохой» INTERNET действительно очень плох, например файл около 180 килобайт пользователь скачивает лишь с вероятностью 0,2.

Вероятность скачивания статических и динамических страниц

Теперь мы в состоянии ответить на вопрос: с какой вероятностью ***любой*** пользователь INTERNET скачает нужный файл (длиной не менее x) ?

Воспользуемся формулой полной вероятности.

Определим все необходимые случайные события и их вероятности.

Событие **A** – пользователь скачал нужный файл (длиной не менее x). Вероятность этого события нужно определить.

Событие H_1 – пользователь имеет возможность скачать нужный файл (длиной не менее x) с первой попытки.

Будем считать, что:

$$P(H_1) = p$$ – вероятность события H_1.

Событие H_2 – пользователь не имеет возможность скачать нужный файл (длиной не менее x) с первой попытки и скачивает его с некоторой вероятностью, которая зависит от x.

В нашем случае события H_1 и H_2 образуют полную систему, поэтому вероятность события H_1 равна:

$$P(H_2) = 1 - P(H_1) = 1 - p$$

Условная вероятность события A при наступлении события H_1 равна 1, то есть:

$$P(A/H_1) = 1$$

Рассмотрим условную вероятность события **A** при наступлении события H_2. Это вероятность скачивания файла пользователем с плохим INTERNET, она зависит от длины файла x. Ранее мы уже определили закон распределения (экспоненциальный) длины скачиваемого файла. Однако нас в данном случае интересует не конкретное значение длины файла x, а значение длины файла не менее x. Поэтому от функции плотности $f(x)$ перейдем к интегральной функции распределения $F(x)$.

Таким образом

$$P(A/H_2) = \int_x^{\infty} f(x)\,dx = 1 - F(x).$$

Интегральная функция распределения дана на рис. 6.

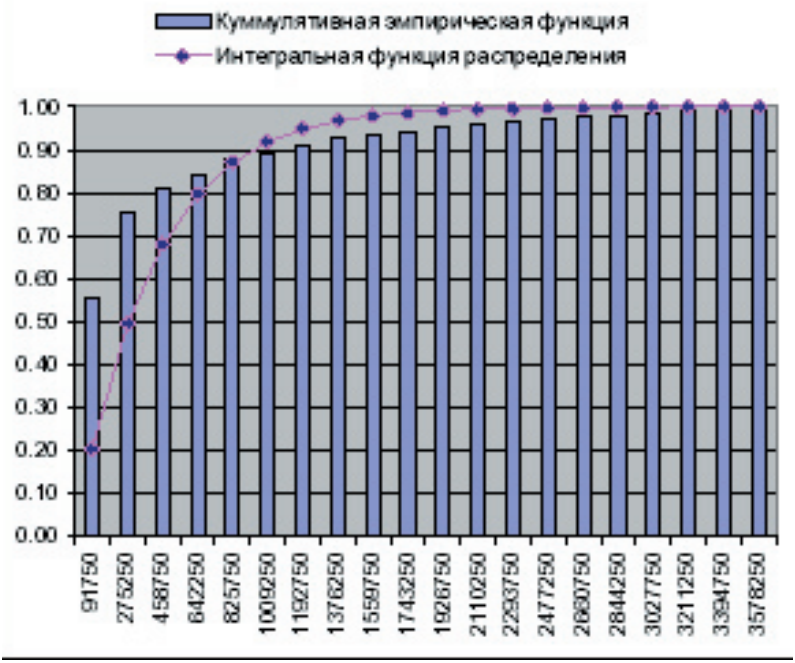


Рис. 6.

Теперь применим формулу полной вероятности:

$$P(A)=P(H_1)\,P(A/H_1)+P(H_2)\,P(A/H_2)$$

или

$$P(A)=p\cdot 1+(1-p)\,(1-F(x)).$$

Для практического применения полученной формулы нам осталось только определить величину p. Для этого воспользуемся классификацией пользователей INTERNET.

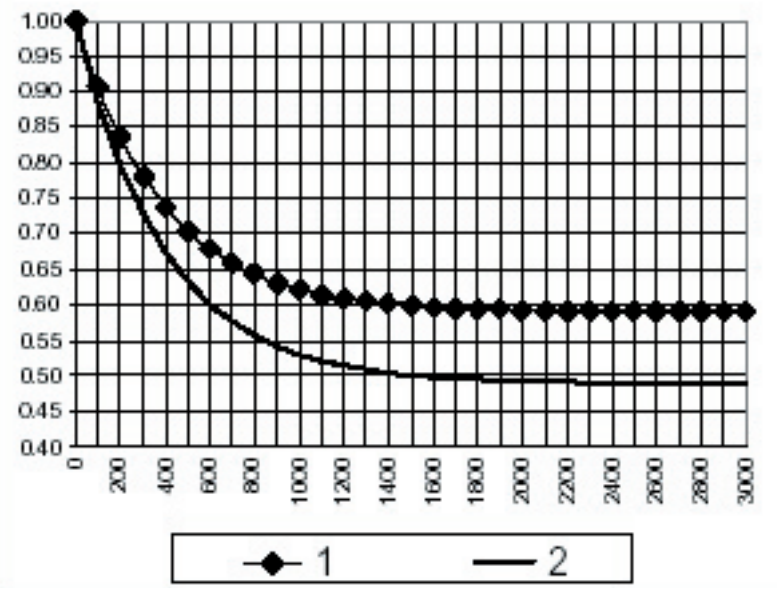
В сети INTERNET принято различать динамические и статические страницы. Нас интересует только тот факт, что статические страницы можно скачивать по частям (с помощью программ докачки), а динамические – нет.

Первый случай

Статические страницы: их наверняка скачают пользователи категорий II (используют программы докачки) и III (хороший INTERNET). В этом случае $p = 0{,}49 + 0{,}10 = 0{,}59; 1 - p = 0{,}41$.

Второй случай

Динамические страницы: их наверняка скачают только пользователи категорий III. Трудно сказать, смогут ли скачать файл пользователи категории II. Выскажем предположение, что INTERNET у них плохой и файл они скачивают точно так же, как пользователи категории I. В этом случае $p = 0{,}49; 1 - p = 0{,}51$.



На рис.7 представлена зависимость вероятности скачивания от размера файла (в килобайтах). Кривая 1 выражает зависимость для процесса скачивания статической страницы, а кривая 2 – для динамической.

Вероятность скачивания файла размером не менее 50 килобайт в обоих случаях приблизительно равна 0,95. Далее кривые расходятся. Для файла не менее 100 КБ вероятности скачивания равны 0,92 (для статических страниц) и 0,87 (для динамических).

Для файла размером не менее 1000 КБ – 0,62 и 0, 53.

Для файлов размером больше 3000 КБ обе кривые проходят параллельно оси X, вероятности скачивания таких файлов приблизительно равны 0,59 и 0,49.

А теперь вернемся к вопросу об «оптимальном» размере файла. Весь анализ, который мы провели, дал лишь «информацию к размышлению». Вывод все равно придется делать человеку. Мы полагаем, что оптимальный размер HTML-страницы (как статической, так и динамической) составляет приблизительно 50 КБ. Этот вывод подтверждается и тем, что такой объем информации, которую программы докачки скачивают с сервера за один раз. Кроме того, примерно на такие же части разбивается информация на «солидных» сайтах (например, в библиотеке Мошкова).

Для других типов файлов (например, электронных учебников или инсталляций программного обеспечения) размер файла уже не имеет значения, так как пользователи с плохим INTERNET практически не смогут получить его без докачки. Однако, если файл слишком большой, то его просто не будут скачивать. Впрочем, это уже предмет совсем другого анализа.

В заключение хотим отметить, что в данной статье мы, естественно, не могли дать описане использованных нами методов математической статистики. Заинтересовавшемуся читателю рекомендуем книгу: Тернер Д. Вероятность, статистика и исследование операций. / Пер. с англ. – М.: Статистика, 1976.

В ней информация изложена очень доступно, без излишней теоретической глубины.